

APPLICATION FOR UNITED STATES LETTERS PATENT

FOR

**Method and System for Generating
Annotations for Video**

Inventors: Rainer Wolfgang Lienhart
Boon-Lock Yeo

Prepared by:
Blakely, Sokoloff, Taylor & Zafman
1279 Oakmead Parkway
Sunnyvale, California 94086
(408) 720-8598

"Express Mail" mailing label number EL 034 146 336 45

Date of Deposit October 29, 1999

I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Assistant Commissioner of Patents and Trademarks, Washington, D.C. 20231.

Lisa Kaiser
(Typed or printed name of person mailing paper or fee)

Lisa Kaiser
(Signature of person mailing paper or fee)

Method and System for Generating Annotations for Video

NOTICE OF COENDING APPLICATION

5 U.S. patent application entitled "METHOD AND APPARATUS FOR
AUTOMATICALLY ABSTRACTING MULTIMEDIA INFORMATION," Application
No. 09/465988, filed on September 10, 1999, assigned to the assignee of the application herein
is referred to below. *Decem 16,*

10 FIELD OF THE INVENTION

The present invention relates to the field of video indexing, retrieval and abstraction. More particularly, the present invention relates to a system and method for on-the-fly video annotations for enabling improved video indexing and retrieval.

15 BACKGROUND OF THE INVENTION

For many years, an increasing number of people own and use video recorders to make video movies that capture their experiences and document their lives. Oddly, most videos are put into a storage box and rarely watched again.

Research is growing in the field of video abstracting. Video abstracting is
20 the processes of taking unedited video footage and combining shorter segments of that
footage into one abstract. Existing automatic video abstracting systems concentrate on
feature films, documentaries or newscasts. Currently, there are generally three systems
that produce videos as abstracts. The first is called video skimming. It aims mainly at
abstracting documentaries and newscasts. Video skimming assumes that the audio track
25 transcript is available. The video and the transcript are then aligned by word spotting.

The audio track of the video skim is constructed by using language analysis (such as the Term Frequency Inverse Document Frequency measure) to identify important words in the transcript. Audio clips around those words are then cut out. Based on detected faces, text, and camera operations, video clips for the video skim are selected from the surrounding frames.

The second system called MoCA Abstracting. MoCA Abstracting was explicitly designed to generate trailers of feature films. The MoCA Abstracting system performs an extensive video analysis of a feature film to segment it into shots or scenes and to determine special events, such as text appearing in the title sequence, close-up shots of main actors, explosions, gunfire, etc. This information is used to select the clips for the video abstract. During the final assembly, ordering and editing rules are presented. Since MoCA Abstracting relies highly on special events such as explosions, gunfire, shot or reverse shot dialogs, and actions that are usually not present in home videos it cannot be used to abstract home video.

The third system by Saarela and Merialdo does not perform any automatic content analysis. Instead they assume that videos have been annotated manually or automatically by descriptors for various properties and relations of audio and video segments. Based on those descriptors the authors try to define "optimal" summaries. They present constraints for video summaries and methods to evaluate the importance of a specific segment.

These existing automatic video abstracting systems concentrate on feature films, documentaries or newscasts. Since raw video footage such as home video is inherently different from all broadcast video, new abstracting principles and algorithms are needed.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates a block diagram of an embodiment of a system utilizing the present invention.

5

Figure 2 illustrates a flow diagram of the steps performed in one embodiment of the present invention.

Figure 3 illustrates a block diagram of adaptive noise cancellation for removing the annotations from the audio signal, according to one embodiment of the present invention.

10

Figure 4 illustrates the method for localizing annotations by signal energy, according to one embodiment of the present invention.

15

Figure 5 illustrates a three-step process for generating enhanced video abstracts, according to one embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Embodiments of the present invention relate to a method and system for generating annotations for video indexing and retrieval. More particularly, in one
5 embodiment the present invention relates to a system and method for on-the-fly video annotations for enabling improved and more selective video abstracts. On-the-fly annotations are verbal notes made by the user contemporaneously with the video footage. Although described with respect to video abstracting, the present invention can be implemented in applications that benefit from indexing and retrieval via annotations. In
10 the following description, for purposes of explanation, specific nomenclature is set forth to provide a thorough understanding of the present invention. For example, the term "video" comprises motion pictures and sound. A "video abstract" denotes a short video consisting of motion pictures and sound. However, it will be apparent to one skilled in the art that these specific details are not required in order to practice the present
15 invention.

The present system allows a user to insert annotations within video footage, use the annotations to create video abstracts and remove the annotations from the original video. The invention may be implemented in multiple embodiments. For
20 example, the video may be annotated "on-the-fly" as described below; alternatively the video may be annotated at a later time. Furthermore, the annotations are not limited to video, but may also be extended to purely audio applications and video obtained by other means. Although the embodiments below implement both hardware components and software components, it will be apparent one skilled in the art that these specific details
25 are not required in order to practice the present invention.

Video Processing System

Referring to FIG. 1, an illustrative embodiment of a system for generating annotations for video indexing and retrieval is shown. The following discussion of the annotation system is described, as it would be implemented when generating a video abstract, by way of example only. One of ordinary skill in the art could implement further embodiments of the annotation system for other purposes.

In one embodiment, the following system utilizes a two-microphone setup with a video recorder to support annotations on-the-fly. The electronic system (100) comprises a processor unit (110), a display device (120), a data storage unit (130), an analog to digital converter (A/D) (140), a video recorder (150), a voice microphone (160), and an environmental microphone (170).

The processor unit (110) acquires a digital video signal (141) from the video recorder (150) via the A/D (140). Voice microphone (160) provides digitized audio annotations via the A/D (140). Environmental microphone (170) provides digitized environmental audio via A/D (140) containing both environmental sounds as well as annotations. The processor unit (110) takes the digital video signal, analyzes it and stores the video together with the digitized voice and environmental audio signals in the data storage (130). When a user requests a new video abstract, the digitized voice signal, environmental audio signal and digitized video signal (141) are retrieved from the data storage (130) in order to create a new video abstract. The video abstract is played back on the display device (120).

The invention would be equally applicable to a system that utilizes a digital video camera with two digital audio tracks or other digital or analog recording

device. Further, the invention could be implemented in a system using less than or more than two microphones.

Still referring to **FIG. 1**, a video recorder with two microphones is setup to support annotations on-the-fly. The voice microphone (160) may be a unidirectional head worn speech input microphone that captures the voice of the cameraman. The output of the voice microphone (160) is stored as the left audio channel of the video recorder (150), while the output of the environmental microphone is stored as the right audio channel of the video recorder (150). Once the video has been transferred to the processor unit (110), the voice channel is transcribed to computer readable text using a voice to text conversion

system. In another embodiment, the audio channels may be switched. The present system can be implemented with existing video recorders, without modification or redesign.

The Processor Unit

FIG. 2 illustrates a flow diagram of the operations performed by the processor unit (110), according to one embodiment of the present invention. In alternative embodiments, the operations shown in **FIG. 2** could be performed by other types of circuitry, such as a DSP, ASIC, or other dedicated circuitry. Moreover, the operations could be a set of instructions stored on a machine readable medium to be executed by the processor unit (110).

At processing block 210 the digital signal (141) is received by a splitter that separates the audio and video channels. At processing block 220 the processor unit (110) removes on-the-fly annotations from the digitized environmental audio. Shot boundaries are generated through video processing block 250 and are used to shorten the duration of each shot. Then, the annotation-free audio signal and shot boundaries are

used to find selected clips of video footage. This process is carried out at shot shortening block 230. Next, the description of a shot's representative clips is stored in the data storage (130). In processing block 240, the digitized annotations (161) are passed on to a speech recognition engine. The resulting transcribed annotations as well as time-stamp annotations are stored in data storage (130).

Annotation Transcription

Annotations may be varying types of free speech. However, the system allows each content descriptive free speech annotation to be preceded by a keyword specifying the kind of annotation and its associated temporal validity. In one embodiment, for the processor unit (110) to transcribe annotations with the speech recognition engine (240), the syntax of the annotations is the following: <keyword>, <temporal validity>, and <free speech text> where keyword is a title, person or object and temporal validity is either of a shot, action, event, day or week. The keyword allows general descriptions to be distinguished from specific descriptions (e.g. naming persons and objects). In practice, object descriptions are most often defaulted if no keyword is present. The temporal validity specifies the temporal extent of the description. By way of example, to label a recording as a one-week vacation video, you would say "TITLE WEEK..." in order to assign the annotation to the vacation video.

In one embodiment, the names SHOT, ACTION, EVENT, DAY and WEEK denote five types of time-based hierarchical shot clusters. An ACTION or level 1 cluster is defined as a sequence of contiguous actions. Shots within a 5-minute duration, form an ACTION cluster. The idea behind this type of cluster is to group shots of high causality. An EVENT or level 2 cluster represents a sequence of contiguous activities within an one-hour duration. EVENT activities usually last longer than actions and span

numerous actions. A DAY or level 3 cluster represents individual days, while a WEEK or level 4 cluster represents seven contiguous DAYS. In one embodiment, if no temporal validity is provided, SHOT or level 0 is assumed. As a result, annotations are not assigned accidentally to unrelated video clips.

5 Editing commands are also supported through Annotations. The first supported command is DELETE. For the processor unit (110) to DELETE sections of video, the syntax of the annotations is the following: DELETE, <time> and <temporal validity> where time is either CURRENT or LAST. For example, a delete command could be "DELETE CURRENT SHOT," or "DELETE LAST EVENT." Another
10 supported editing command is for RATING. For the processor unit (110) to rate the importance of contents, the syntax of the annotations may be the following: RATING, <time><temporal validity> as <rating> where rating is either SECONDARY, DEFAULT, IMPORTANT or HIGHLIGHT. In other embodiments, alternative editing commands can be supported without departing from the present invention.

15 **Enhancing Video Abstracts**

Referring to FIG.2, which illustrates a flow diagram of the processor unit's (110) functions, annotations enable semantic indexing and retrieval and are useful in diverse applications. Basically, annotations can be exploited in many ways to enhance
20 the video extracts in the video processing block 250. For example, an annotation can be used to select the source video material that is to be abstracted allowing a viewer to watch thematic abstracts. Hereto, the user specifies a text query. All shot clusters whose annotations meet the query are taken as the source video material. For instance, assume that a user wants to view an abstract of all birthday parties of his daughter Nadine. By
25 specifying "Nadine AND birthday" as the query, all clusters containing both words in

their associated annotation(s) will be selected as the source video set. An abstract is then created based on the source video set. The user may also specify that the cluster level of clusters retrieved by some query words should be increased for certain query terms with respect to the actual cluster level to which the query term is assigned. For example, if
5 only one shot is annotated "Nadine's birthday," but more than one shot has been recorded, you can request to retrieve the level 2 cluster associated with "Nadine's birthday" which may generate all the shots of the birthday.

Secondly, annotations can be used to enhance the visual presentation of video abstracts. Hereto, short TITLE annotations may be shown as centered text title
10 introducing the subsequent video clips. Annotations might also be added as small scrolling text banners at the bottom or top of the video frames. PEOPLE annotations can be used to append a closing sequence to the video abstract.

Finally, shots can be cut down to the part(s) with descriptive annotations, and annotation-free shots may be discarded. This abstraction rule is motivated by the fact
15 that the important objects, persons, or actions often come along with annotations. Nevertheless, additional clips may either be discarded or added in order to meet the target duration of the requested abstract.

Removing Annotations

20 Referring now to **FIG. 3** which illustrates a block diagram of adaptive noise cancellation for removing the annotations from the audio signal in two steps according to one embodiment. In the first step, an annotation detector examines the audio signal in a window of a one second duration and counts the number of amplitude values
above a threshold. In alternate embodiments the annotation detector may examine the
25 audio signal in windows less than or greater than a one second duration. If the count

exceeds a certain number, the window is declared to contain annotations. Then each annotation window is expanded by half a second to the left and a second to the right in order not to cut off the beginning and ending of the annotation. Overlapping annotation windows are merged into one large annotation's occurrence. The processor unit (110)

5 removes annotations from the environmental audio signal. The least-mean-square (LMS) algorithm developed by Widrow and Hoff may be used to remove the annotations. In other embodiments, annotations may be removed by other methods or algorithms. In LMS filtering a transversal filter (310) estimates the noise signal (e.g. the signal of annotations) in the environmental sound. The difference between the environmental
10 sound and the estimate of annotations is used to calculate the estimation error. Based on the estimation error, the adaptive filter (320) then adjusts the tap weights to reduce the mean-square value of the estimation error.

One potential problem of the LMS algorithm is its slow adjustment rate to the real tap weights of a wide-sense stationary signal as well as to system changes of the
15 underlying signal. Referring to **FIG. 4**, in order to improve the quality of removal at the beginning of each annotation segment, noise cancellation is not only performed forward in time but also backward over all detected annotation ranges (410). The backward noise cancellation is used to estimate the annotation-free audio signal for the first half of each annotation segment (430). While the forward noise cancellation is used to estimate the
20 annotation-free audio signal for the second half of each annotation segment (440). This scheme improves the beginning of annotation segments.

SKD
72

Creating Video Abstracts

Varying methods for creating video abstracts may be implemented with
25 the present system. By way of example, a video abstract may be obtained from the

ST
AZ

methods disclosed in co-pending provisional U.S. patent application entitled "METHOD AND APPARATUS FOR AUTOMATICALLY ABSTRACTING MULTIMEDIA INFORMATION." Application No.____, filed on September 10, 1999, assigned to the assignee of the application herein.

5 In one embodiment, video abstracts are created in 3 steps. First a one-time processing step analyzes the video in order to determine Shot Boundaries (511), Time and Date of Recording (512), Hierarchical Shot Clusters (513), and Short Representative Clips (514).

Shot Boundaries (511) are determined based upon how the video was
10 acquired, whether with a digital video camera or analog video recorder. If a digital video camera is used, the Shot Boundaries (511) are calculated based upon the time and date of the recording stored with the frames. In one embodiment, time increments of more than one second between contiguous frames marks a Shot Boundary (511). With analog video cameras, video is first digitized and Shot Boundaries (511) are then determined.

15 Time and Date of Recording (512) are also determined based upon how the video was acquired, whether with a digital video camera or analog video recorder. If a digital video camera is used, the Time and Date of Recording (512) can be read on the digital video stream. In the case of an analog video recorder, a superimposed date and time stamp created on the film when the video was acquired can be extracted. By way of
20 example, this information may be extracted from the video by adjusting the text segmentation and text recognition algorithms to the specific characteristics of time and date information in the videos.

Hierarchical Shot Clusters (513) are constructed based upon the Time and Date of Recording (512), thus generating a four-level Hierarchy of Shot Clusters (513).
25 The Shot Clusters (513) are determined by temporal distance between contiguous shots.

Individual shots represent a level 0 cluster, while levels 1, 2, 3, and 4 represent a sequence of contiguous actions, activities, individual days and weeks.

In one embodiment, Short Representative Clips (514) for shots are created when a shot is reduced to a relevant part based on audio features. By way of example, the shot may be reduced to approximately 10 seconds of video. Finally, these representative clips that are to be part of the video abstract are selected and compiled into the final video abstract by inserting edits and title sequences (530). Clip selection and edit decisions are based upon heuristic rules. For example, shots of different days are always spliced together by a fade out or fade in sequence making the temporal gap explicitly visible.

The foregoing has described a method and system for creating video abstracts. More particularly, the present invention relates to a system and method for on-the-fly video annotations for enabling improved and more selective video abstracts. It is contemplated that changes and modifications may be made by one of ordinary skill in the art, to the materials and arrangements of elements of the present invention without departing from the scope of the invention.

SUMMARY OF THE INVENTION

The present invention provides a method and system for processing video signals. The system receives a video signal, a first audio signal containing an annotation and a second audio signal containing environmental sounds corresponding to the video
5 signal. In one embodiment the system generates searchable annotations corresponding to the video and second audio signals via the first audio signal.

SECRET - REF ID: A66007